FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Techniques to deal with imbalanced data in multi-class problems: A review of existing methods

Vitor Miguel Saraiva Esteves

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: João Pedro Mendes Moreira

February 3, 2020

Techniques to deal with imbalanced data in multi-class problems: A review of existing methods

Vitor Miguel Saraiva Esteves

Mestrado Integrado em Engenharia Informática e Computação

February 3, 2020

Abstract

Imbalanced learning is one of the most challenging classification problems in the field of machine learning and has been receiving increased attention over the years due to the necessity of handling real world information which is usually skewed. This effect occurs when one of the classes has a bigger number of examples compared to the rest. When we attempt to classify data in said conditions, machine learning algorithms will be able to correctly identify the majority class examples but will most likely fail when attempting to identify minority class examples, which often end up representing the most valuable information. Several surveys were published in the last few years about approaches to solve the problem. Adding to this issue, when we have a multi-class scenario where the examples to be classified can fall into more than two classes, the accuracy of applied techniques plummets due to their inability to deal with the issue. In the latest years, various techniques were proposed to deal with the matter, and they usually do so by converting the problem into subsets of two-class problems that can be solved by common classifiers. This process is called class decomposition. However, recent studies show that it can cause class overlapping and loss of valuable information. In this study, we address the topic first, by identifying algorithms that deal with multi-class imbalance without using class decomposition and categorizing them based on their approach. Then, we proceed to benchmark the two latest state-of-the-art ensemble algorithms, MBBR and SOUPBagging against each other and see how they perform in 9 real life datasets. The results showcase their ability to handle multi-class imbalance with high accuracy for classifying skewed data. Lastly, a possible future direction in the field is briefly discussed.

Keywords: Multi-class imbalance learning, Sampling techniques, Cost-sensitivity frameworks, Algorithm-level approaches, Ensemble-Based learning, Classification, Evaluation Metrics

Resumo

O tratamento de dados desbalanceados constitui um dos problemas mais complicados de resolver no campo do *machine learning* e tem vindo a receber maior atenção ao longo dos anos devido à necessidade de lidar com informações do mundo real que geralmente apresenta esta caracteristica. O desbalanceamento ocorre quando uma das classes possui um número maior de exemplos em comparação com o restante e, quando tentamos classificar os dados nessas condições, os algoritmos de *machine learning*, apesar de conseguirem identificar corretamente os exemplos das classes maioritárias, normalmente são incapazes de classificar os de classe minoritária. Cria-se então um problema pois, por norma, a informação destas classes é mais valiosa do que das restantes. Vários artigos foram publicadas nos últimos anos sobre abordagens para resolver o problema. Para além disso, se considerarmos um cenário de classes múltiplas em que o resultado na classificação não recai apenas em duas classes, a precisão das técnicas diminui consideravelmente devido à sua incapacidade de lidar com o problema. Nos últimos anos, várias técnicas foram propostas para lidar com este assunto, e geralmente a abordagem escolhida recai sob a forma de divisão do problema em subconjuntos de problemas de duas classes que podem ser resolvidos por classificadores comuns. A este processo chama-se decomposição de classes. No entanto, estudos recentes mostram que isso pode causar sobreposição de classe e perda de informações valiosas. Nesta dissertação o tópico é explorado, primeiro, através da identificação de algoritmos que lidem com problemas de dados desbalanceados e classe múltipla sem que utilizem decomposição de classes e em seguida através da categorização dos mesmos com base na abordagem seguida. Procede-se também à avaliação da qualidade duas técnicas de *ensemble learning*, de nome, *MBBR* e *SOUPBagging* através de uma comparação na classificação de 9 datasets com informações do mundo real. Os resultados obtidos demonstram alta precisão e comprovam a habilidade destes algoritmos no tratamento de problemas de dados desbalanceados e classe múltipla. Por fim, é feita uma análise onde se discutem possíveis caminhos de futuro para o tópico abordado.

Palavras-chave: Problema de dados desbalanceados de classe múltipla, Técnicas de balanceamento de dados, *Cost-sensitivity frameworks*, Abordagens a nivel algorítmico, Aprendizagem *Ensemble-Based*, Problemas de classificação, Métricas de avaliação

Acknowledgements

To my parents, Vitor and Paula, for the unconditional love and support throughout the years that made me who I am today and to my girlfriend Catarina for the patience and strength that helped me push through. To all my friends who listened to me when I was in a bad mood including professor João Moreira, who helped me reach my academic goals.
To all of you who in some way or another helped me grow, thank you.

Vitor Miguel Saraiva Esteves

"Opportunities multiply as they are seized."

General Sun Tzu

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

Automating the process of event prediction has been crucial to the development of society and technology, especially with the rampant growth of user activity on the Internet which generates a flow of information too large to be analyzed by a human but of extreme value to do business or research with.[PMS18] In order to solve this problem, machine learning techniques were developed. In the field of machine learning, a prediction problem is often designed as either a problem of classification or regression. By definition, predictive modeling is "the attempt to build a model to find an unknown function Y = f(X1,X2,... ,Xp), based on a training sample with examples regarding this function. The type of value of the Y function can either be nominal, and in this case we are facing a classification problem, or quantitative, which means we face a regression task."[BTR16] Both of these instances are included in a specific category of learning processes, called supervised learning. This report will focus on the first type, more specifically in a subset of classification problems called multi-class. A multi-class classification problem can be defined as the attempt to classify data where the instances may be a part of one of three or more classes. An easier way to understand this concept is thinking about a hypothetical attempt to classify fruits based on their characteristics. Our dependent or target variable can be of several types, such as, apples, bananas, strawberries or mangoes. In a classification problem, when one or more classes outnumbers (or is/are outnumbered by) the rest it is said that the class(es) is/are imbalanced. This means that these classes have more examples compared to others[AAZ+18] which will most of the times lead to worse predictions because the more commonly found machine learning algorithms are built disregarding the balance between classes. We can find this adversity in several real life examples such as Plankton Image Classification[DWT+18], Crime Prediction in Smart Cities[PPEP19], Protein Data Classification[S+14], Weld Flaw Classification[Lia08], Sentiment Analysis[KMC17]. That said, the main objective of this dissertation is reviewing the existing techniques used to tackle the issue of imbalanced data in multi-class problems.

1.1 Motivation

Dealing with class imbalance is a difficult task that is researched daily due to the real world implications of advances in the field. Combining class imbalance and multi-class into the same problem has strong negative effects on the capacity of common classifiers and degrades predictions. Usually, to attempt to lower said effects, conversion into subsets of binary problems is the easiest solution. However, not only computational costs for said approach are high, reports[KMC17][WY12a] show that this is not an effective technique due to loss of information and over generalization that leads to bad performance. This means that there should be more research into investigating techniques that deal with multi-class imbalance without class decomposition. Besides, regarding current research, there is a need for agglomeration since information is scattered and hardly understandable, which creates even more difficulties.

1.2 Objectives

The main objective of this study is identifying and categorizing algorithms that were developed to deal with multi-class imbalance and afterwards, from the outcome of said process, draw conclusions that could lead to improvements in the area. To do so, an in-depth study of the state of the art regarding the topic should be conducted in order to be able to identify potential key points that could lead to a breakthrough in the field.

1.3 Document Structure explained

In section 2, a brief explanation will be given about some key concepts related with machine learning such as the difference between supervised and unsupervised learning, binary versus multi-class classification and other relevant details. In section 3, there will be a discussion about the methodologies to address the topic in question and also a description of some specific multi-class problems. Section 4, represents a survey on multi-class methods that don't use class decomposition and fit into the following categories, respectively: algorithm-level techniques, data-level techniques, cost-sensitive frameworks and ensemble learning algorithms. A brief explanation about class decomposition schemes is also provided for added context. Evaluation metrics are also discussed. In section 5 a benchmark comparison between two state of the art sampling methods and two ensemble learning methods is conducted and results discussed. Finally, section 6 conclusions from the conducted study are provided and thoughts on future directions on the topic are shared.

Chapter 2

Key Concepts

In this section some key concepts about machine learning will be introduced regarding multi-class and imbalance problems in order to clarify future references that will be made on the course of this study.

2.1 Supervised vs Unsupervised Learning

Supervised learning is a labeled process where we attempt to classify data by mapping from input to output labels, or regression, by mapping input to continuous output with the objective of finding similarities or structure in the input examples that allow us to correctly identify output data. Some of the most common are logistic regression, random forests, support vector machines and artificial neural networks. On the other hand, common unsupervised learning techniques such as clustering and representation attempt to learn the structure of our data without the usage of labels. The most common algorithm is k-means clustering.[sup19]

Figure 2.1: Comparison between supervised and unsupervised learning objectives [sup19]

2.2 Imbalanced Datasets

A dataset can be considered imbalanced, if one or more of the classes have a larger proportion of examples comparatively with the others. These classes are called majority classes and the others are called minority classes.[goo19]

In the example of credit card fraud data, the existence of reported fraud has a very small amount of instances comparatively to no fraud and so, it represents our minority class

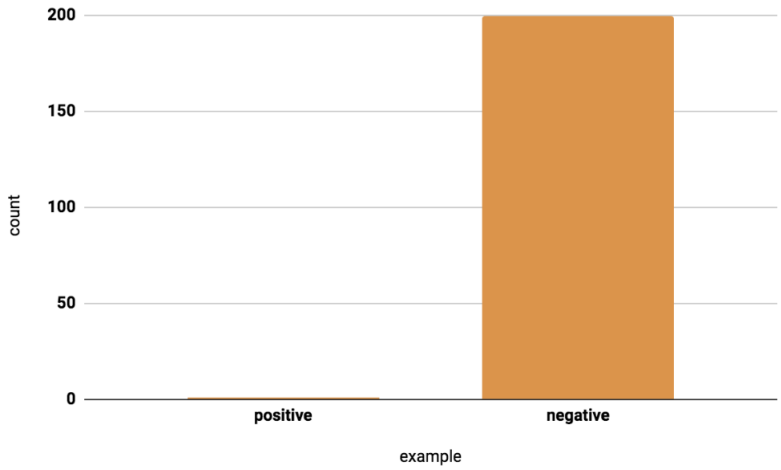
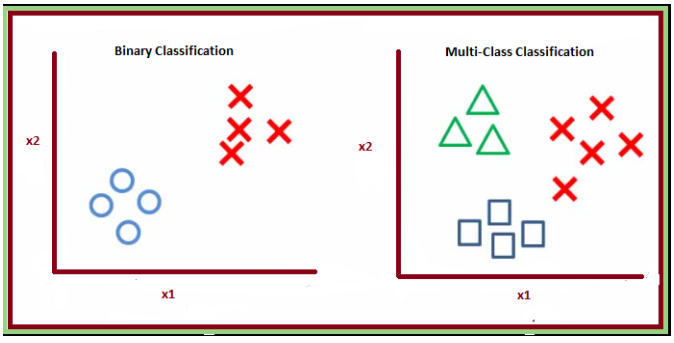


Figure 2.2: Credit card fraud dataset representation. [goo19]

2.3 Binary vs Multi-class classification

A multi-class problem is a classification problem where the instances of data to classify fall into one of three or more classes. Binary classification, as the name implies, is a classification problem where the instances of data to be classified can fall into exactly two classes. The most common solution to deal with multi-class problems, since they tend to be harder to predict, is transforming them into several binary problems.

Figure 2.3: Binary versus multi-class classification [bin19]



Chapter 3

Methodology

Even though the topic sparked interest in recent years[Lan19][ZBX+19][BZ18][HYS+17], there is still a lack of systematic and organized research and most of the effort made drives toward problem-oriented solutions which convert multi-class problems into binary. That said, we will attempt to categorize algorithms that deal specifically with multi-class scenarios without the necessity for class decomposition, but also try to understand the differences between both approaches, describing their most commonly used techniques and how well they perform for the same problems.

3.1 Ambiguity on the taxonomy and application

The first problem we delve into is the inability to find a robust taxonomy to categorize techniques and strategies to deal with the problem of class imbalance. Haixiang et al.[HYS+17] distinguishes between two basic strategies which are preprocessing and cost-sensitive learning. These strategies are further integrated into classification models which will then be divided into ensemble-based classifiers and algorithm modified classifiers. Even though it is uncommon to apply the second strategy on multi-class problems, a cost-sensitive algorithm which addresses the issue without class decomposition was proposed lately.[SKW06] It focuses on finding an appropriate cost matrix with multiple classes and then the costs are introduced into the algorithm. As another perspective, we see Yasir Arafat et al., considers the distinction of three strategies to tackle the issue, such as sampling techniques (a specific type of pre-processing), which include techniques of over and undersampling, cost-sensitive learning methods, that take the misclassification costs into consideration during the learning process in order to minimize total costs and ensemble-based methods, that combine different classifiers to form a strong classifier that attempts to identify new examples with high prediction accuracy. Meanwhile, R.Cruz et al.[CSSC18] and M.Galar et al.[GFB+11] go further and clearly create a distinction between four strategies: algorithm-level approaches, which

modify the usual learning methods so they take into consideration the imbalance between different classes, data-level approaches, which include the sampling pre-processing techniques that are used to diminish the impact of the imbalanced class during the learning process, cost-sensitive learning frameworks, that combine both data-level and algorithm-level approaches by assigning different costs to every class based on the distribution of examples and by applying modified learning algorithms to these classes, and last, ensemble-based approaches, which combine any of the previous techniques, especially preprocessing, with an ensemble-based algorithm. That said, we can identify both similarities and differences in the development process of each researcher, which might lead to ambiguity when attempting to address the topic thus creating the need to establish a clearer definition for new researchers moving forward. The second problem we face is the categorization of said strategies when applied to multi-class classification imbalance problems instead of the more commonly solvable binary tasks. Most existing solutions only use class decomposition schemes to handle multi-class and work with two-class imbalance techniques to handle each imbalanced binary subtask.[WY12b] In two of the examples shown before, both the protein-fold problem and the weld flaw classification make use of techniques such as One-Against-All (OAA) and One-Against-One (OAO) to decompose classes into binary and then build learners in order to improve the representation of the minority examples. A further explanation on how these and many other schemes work will be given in the next chapters when we attempt to categorize algorithms applied to multi-class classification imbalance based on the four strategies thought process used to solve common binary class scenarios. Lastly, we will address the assessment metrics used in class imbalance learning since the most common ones such as F-measure, recall and G-mean were originally designed for two-class problems.

3.2 Multi-class specific problems

When dealing with class imbalance in a multi-class problem we also face different irregularities not found in two class problems, such as, the possibility of having not only one majority class but several and also the opposite, one majority class and several minority classes. Shuo Wang et al.[WY12b] attempts to understand the robustness of sampling techniques when applied to multi-class problems by applying AdaBoost coupled with both oversampling and undersampling techniques to artificial cases, by replicating examples from all minority classes randomly until each one of them has the same size as the majority class or by eliminating examples from the majority classes randomly until each of them has the same size as the minority class before training starts. It is then concluded that oversampling does not help in either multi minority and multi majority cases because it causes overfitting of the minority-class. Undersampling techniques on the other hand, in the multi-minority case, can be sensitive to the class number while in the multi majority case, there is a high risk of sacrificing too much majority-class performance. If we think about the problem of multi class imbalance itself and not just the techniques applied, the multi majority situation seems to be more difficult than multi minority.

Chapter 4

Survey on multi-class imbalance techniques

This chapter starts by giving a brief explanation on class decomposition most used techniques just to clarify terms that will be referred throughout the rest of the survey. In the following sections, the discussed taxonomy is applied to categorize algorithms based on their approach and a description about how do they work is given. Results are discussed individually for each method.

Class decomposition techniques

Class decomposition schemes are the most frequently used solution to deal with multi-class problems. Even in the context of imbalanced data, the approach is transforming the problem into a two-class imbalanced subtask that can be learned by binary classifier techniques. In the protein classification example given before, two schemes are used, one-against-all (OAA) and one-against-one (OAO), to subdivide the problem allowing for better results. Meanwhile, in the weld flaw classification example, resampling techniques are used cooperatively with OAA. These two methods are the most popular when dealing with class decomposition. OAA was first discussed as a technique by Clark and Boswell (1991)[CB91] to improve the performance of CN2[CN89], a rule induction algorithm but the name one-against-all was first given by Fuernkranz (2006)[BFH06]. It labels each class as positive and all the other classes as negative applying the same method for all classes and then trains a classification model. The result is multiple binary classification models that can be compared in order to find the best. In OAO, a subset from the dataset is selected containing only examples from two classes to make a pairwise comparison by training a classifier for each pair of classes. During prediction it will be used conjunctively with all the other pairwise comparisons in order to predict a new example. A method combining both schemes, All-and-One (A&O)[GPOB06] was also proposed as an attempt to combat the techniques' downsides. First, it uses OAA to find the best two predictions and then a pairwise comparison is made between them using OAO to deliver a final prediction. The main problem of class decomposition techniques is the inability to integrate all the information coming from the trained classifiers and the lack of

good results when dealing with imbalanced data, especially for OAA. Meanwhile, OAO is prone to overfitting[LDCG08] when datasets have a small number of examples. More recently, Murphey et. al designed a decomposition method that specifically deals with class imbalance, One-Against-Higher-Order (OAHO).[MWOF07] The first step in this method is sorting classes by number of examples from largest to smallest. Starting from the first class until the last, the current class will be deemed positive and the rest will be marked as negative and a binary classifier will be trained. If the prediction result is equal to the sample that is our final prediction, else it will move onto lower ranks until it finds a match.

4.1 Data-level approaches

Data-level approaches are related with the techniques applied to raw data to modify its distribution in order to minimize the effect the majority class has on the outcome of a learning algorithm, which often yields poor predictions in cases of imbalance. This happens during the data-preprocessing phase, before building the model, in order to obtain better input examples by having algorithms operating the datasets.

Preprocessing techniques used in multi class imbalance problems are either about resampling data or feature selection. In this section an explanation of what both methods are, how their most common algorithms work when applied to multi-class specific problems, and an insight on class decomposition schemes will be given, since conversion of multi-class to binary is one of the most commonly used techniques.

4.1.1 Resampling

Resampling, by definition, is the method applied to data to respectively select or generate a specific amount of examples from the majority or minority classes in order to rebalance them and diminish the impact of imbalance. The main methods used to resample data can be classified into three different groups which are, undersampling, oversampling and hybrid methods. Undersampling methods create a subset of the original dataset by eliminating some of the examples of the majority class. On the contrary, Oversampling methods create a superset of the original data-set by replicating some of the examples of the minority class or generating new ones from the original minority class instances. Mixed or hybrid methods combine the two previous methods, eliminating some of the examples before or after resampling, in order to reduce overfitting.[FLG+13] The most frequently used algorithms in any of these methods fit into two main categories, which are cluster-based or distance-based.[HYS+17]

4.1.1.1 Oversampling

The most simple technique, random oversampling, replicates examples from the minority class randomly as an attempt to deal with the data skewness. Even though it doesn't create new infor-

mation since the copies are from existing entries, it increases the possibility of model overfitting. That said, this subsection contains a categorized list with a brief explanation of some oversampling algorithms that are often utilized in multi class scenarios with no class and attempt to avoid overfitting.

Distance Based Algorithms

SMOTE and variants

Proposed in 2002 by Chawla et al.,[CBH] the Synthetic Minority Oversampling Technique (SMOTE) is a method where synthetic samples are generated from the minority class based on their examples similarities' and it is one of most commonly used techniques when attempting to resample data either on one class or multi-class scenarios. Depending on the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. This can also cause a problem of overgeneralization because the majority class is not taken into account. Also, since the generated synthetic samples also disregard the majority class, it can lead to overlapping between classes especially in multi-class cases.[PA04]
To overcome said flaws and, according to a publication from earlier this year, more than 85 variants[Kov19] of SMOTE have been published since its development, the most notable being:

-Safe-level SMOTE[BSL09] assigns each positive instance its safe level before generating synthetic instances. Each synthetic instance is positioned closer to the largest safe level so all synthetic instances are generated only in safe regions. The safe level is the number of positive instances in k nearest neighbours. If the safe level of an instance is close to 0, the instance is nearly noise. If it is close to k, the instance is considered safe.

-Borderline SMOTE[HWM05] is a technique where only the borderline examples of the minority class are over-sampled. First, these borderline examples should be accurately identified and then, synthetic entries are generated from them and added to the original training set. Even though this technique performs better in two-class examples, it still suffers from the same original techniques' problems when dealing with multi-class due to the augmented risk of overlap and overgeneralization.

-ADASYN[HBGL08] is an algorithm that uses density distribution as a criterion to automatically decide the number of synthetic samples that need to be generated for each minority data

example. We can consider this criterion as a measurement of the distribution of weights for different minority class examples according to their level of difficulty in learning. ADASYN will adaptively generate synthetic data samples for the minority class to reduce the bias introduced by the imbalanced data distribution and shift the classifier decision boundary to be more focused on those difficult to learn examples, therefore improving learning performance over SMOTE.

-KSMOTE[PS12] is a combination of two techniques, K-means and SMOTE. The K-means algorithm is used for splitting the dataset into two clusters. If the majority class and the remaining classes in each cluster have imbalanced ratio higher than a threshold, SMOTE algorithm is applied to re-balance the data distribution in the cluster. Moreover, random undersampling is used to balance the class distribution.

MDO/AMDO

MDO[AH15] is an oversampling technique which generates synthetic samples for each minority class that have the same Mahalanobis distance between the considered class mean and the minority class candidate example. By generating these samples in dense areas of the feature space and considering suitable samples in each minority class it will reduce the risk of overlapping between different class regions without the need for class decomposition. Even though it outperforms well-known methods, it might experience some issues such as the possibility of generating excessive or unrealistic samples and, its biggest liability lies on the fact that it can only be applied to datasets with numeric attributes. This lead to the development of an extended technique called adaptive Mahalanobis distance oversampling (AMDO)[YKZZ17]. By using Heterogeneous Value Difference Metric (HVDM)[WM97] and Generalized Singular Value Decomposition (GSVD)[CKSLS14] the initial algorithm is converted to accept mixed-type attributes, which solves the main issue. Over-generation is dealt with by the means of a partly balanced resampling scheme and sample impossibility generation is diminished. The extended version will outperform MDO both in mixed-type/numeric datasets and in multi class scenarios.

MC-RBO

Developed to specifically handle multi-class imbalanced problems, Multi-Class Radial-Based Oversampling (MC-RBO)[KKW19] is an algorithm that attempts to improve upon most of the well-known binary oversampling techniques' (including its predecessor's, B-RBO[KKW17]) flaws when dealing with atypical data distributions and overlapping classes. The proposed method is an iterative procedure where each class of the dataset will be oversampled individually: The first step is sorting out classes by number of examples from most to least. Then, the one on the top, the majority class, is put aside and for all the other classes a combination of majority examples is constructed. Then, oversampling is performed with B-RBO by using each of the classes already considered before as minority and the constructed class as the majority iteratively. In comparison

with using class decomposition to adapt the original method, MC-RBO displays two clear advantages which are: Reduced computational costs, since the combined majority examples are less than the entire collection of samples. Assigned equal weights to all classes in the combined collection of majority examples, since all of the classes have an equal number of samples. The same could not be done by any class decomposition technique, since the classes with a higher number of samples would dominate the rest. The results also point towards MC-RBO having a better performance than the most used class decomposition techniques such as OAA or OAO and other multi-class sampling techniques discussed previously such as MDO, SMOTE and ADASYN. Since several algorithms were compared in this publication, it is also noteworthy that most of the methods that do not use class decomposition perform better than their counterparts which inevitably showcases the inferior performance of the second ones due to phenomena such as class overlap, or the inability to detect outliers since there is no longer a complex view on initial class' structures.

Cluster-Based algorithms

CBOS

Cluster Based Oversampling (CBOS)[SD18] is a technique which uses euclidean distance in k-Means clustering to generate the cluster centroid. A distance normalization technique is used to generate the number of new data samples per existing minority class sample to be created. This algorithm is best suited for high dimensional data settings because with less number of attributes there is a possibility that the samples generated are too similar, which may eventually lead to overfitting. There is a clear distinction between this algorithm and more commonly used SMOTE because of the way this algorithm decides how many new samples are to be generated for each existing minority class sample. The idea is that the more distant a sample point is from its cluster centroid, the lower the number of new points of this sample will be generated; and the less distant a sample point is from cluster centroid, the more is the number of new samples associated with this sample to be generated. Clustering also helps to add the spatial structure of minority class into the new generated data samples and the non-dependence of this algorithm on the clustering technique used makes it more stable preventing any change in the learning performance of the majority class. The way the algorithm decides on the number of new samples for each original instance makes sure that the more important samples have higher representation than far lying samples in the new balanced minority class space.

TRIM

TRIM[PRW16] is a cluster based minority oversampling algorithm whose design is based on SMOTE and that attempts at solving the overgeneralization problem faced by this technique. TRIM searches for precise minority class regions and iteratively filters out irrelevant majority data and then outputs multiple subsets of minority class examples generating synthetic data samples via SMOTE. The first step consists of splitting the dataset into a number of small clusters

based on the clusters' precision and generalization. Even though there might be a chance for minority and majority classes to be mixed together in a cluster, most of the majority class examples will be pruned. Then, two minority class samples from different clusters are connected based on a proximity criterion and the output of this process is the set of the minority class pairs which are used to create new synthetic data.

4.1.1.2 Undersampling

Unlike oversampling, random undersampling does not cause overfitting and is in fact one of the most effective methods of dealing with data imbalance. The downside is the possible loss of valuable information. A brief explanation on some state of the art undersampling techniques used in multi-class problems will be conducted in this subsection.

Cluster-Based algorithms

DSUS

The proposal of the Diversified Sensitivity-Based Undersampling (DSUS)[NHY+15] technique is tied to the lack of consideration of informative samples in the datasets and disregard for information distribution which inevitably may lead to class overlapping when dealing with imbalanced data by the most used undersampling techniques such as Random Undersampling (RUS). Thus said, this algorithm works by first dividing the majority class into clusters by using k-means. From each of those clusters a representative sample is selected. Then, by utilizing stochastic sensitivity measures (SM)[Irv92] undersampling is performed via sample selection. Lastly, the samples selected by SM are used to train a Radial Basis function neural network (RFBNN). Even though this method was originally created to deal with binary classes, an extension was made, and when facing more, the one with the least representative is chosen as the minority class, and samples of the other are treated as the majority class. Results showed it to perform better than other undersampling techniques, including RUS, in this scenario.

Distance-Based algorithms

RBU

Radial-Based Undersampling (RBU) is an algorithm based on one of the methods discussed before, RBO, which attempts to explore the possibility of using the same concept of mutual class potential in an undersampling approach to solve class imbalance.

4.1.1.3 Hybrid sampling

In this subsection a brief explanation on hybrid methods that combine oversampling and undersampling techniques will be given.

SCUT

SCUT[AVP15] employs a cluster-based technique, the Expectation Maximization (EM) algorithm, for undersampling and SMOTE for oversampling to solve multi-class imbalance problems without the need for class decomposition. The algorithm works by first splitting the dataset into n parts where n is the number of classes and by calculating the mean m of the number of instances of all the classes. Depending on the number of instances of each class, a different process will be followed. - For all classes that have a number of instances less than the mean m, oversampling is performed in order to obtain a number of instances equal to the mean. The sampling percentage used for SMOTE is calculated such that the number of instances in the class after oversampling is equal to m. -For all classes that have a number of instances greater than the mean m, undersampling is conducted to obtain a number of instances equal to the mean. The EM technique is used to discover the clusters within each class. For each cluster within the current class, instances are randomly selected such that the total number of instances from all the clusters is equal to m. Therefore, an attempt is made to select the instances as uniformly as possible. The selected instances are combined in order to obtain a value of instances equal to m. - For all classes for which the number of instances is equal to the mean m, nothing is done. Finally, all the classes are merged together in order to obtain a dataset D', where all the classes have m instances. Classification may be performed on D' using an appropriate classifier.

C-MIEN

Clustering with sampling for multi-class Imbalanced classification using ensemble (C-MIEN)[PS14] is an algorithm which uses both clustering and hybrid sampling techniques. The first step of this method consists of applying k-means clustering to divide the dataset into two clusters whose elements share similarities. The next step is applying both undersampling and oversampling techniques to both clusters independently in order to reduce overfitting and prevent loss of information. Lastly, to classify data, two experiences are made: for the first one (S1), a classifier is trained in each cluster; for the other (S2), base classifiers are trained independently on the subsets of the training set in each cluster. After, both clusters are combined through majority vote in order to attain a prediction. C-MIEN has five different experimental setups (C-MIEN1 to C-MIEN5) which used different imbalance ratios and rebalance processes. C-MIEN5 will be discussed later, since it uses ensemble. The experimental setup is described as follows.

Experimental setup for studying the behavior of C-MIEN

Name	Re-clustering process Fig. 1(a)	Re-balancing process Fig. 1(b)		Imbalance ratios	Ensemble
		S_1	S_2		
C-MIEN1	✓	✓	✓	1.1 to 2.0	✗
C-MIEN2	✓	✗	✗	✗	✗
C-MIEN3	✓	✓	✗	1.5	✗
C-MIEN4	✓	✓	✓	1.5	✗
C-MIEN5	✓	✓	✓	1.5	✓

The results obtained by the first four methods are then compared with different algorithms and some decomposition techniques: Decision trees (C4.5), OAO, OAA, Decision Trees with

Resampling (RC4.5), One-Against-One with Resampling (ROAO) and One-Against-All with Resampling (ROAA). The results obtained by the C-MIEN5 ensemble technique are compared with the algorithms: Bagging with Decision tree (BC4.5), Bagging with One-Against-One (BOAO), Bagging with One-Against-All (BOAA), Bagging of Decision tree with Resampling (BRC4.5), Bagging of One-Against-One with Resampling (BROAO), Bagging of One-Against-All with Resampling (BROAA), AdaboostM1 and MultiBoosting. From the obtained results for the first four techniques it is concluded that:

-For F-measure, C-MIEN4 outperforms the other methods in most datasets while also obtaining the best rank in the Nemenyi test while C-MIEN2 and C-MIEN3 perform better in two out of the seven datasets. Average F-measure of all C-MIEN methods also outperform ROAA which ranks 4th in best performance.
-For G-mean, C-MIEN4 again outperforms its competition in most of the datasets and ranks best again in the Nemenyi test, while C-MIEN2 and C-MIEN3 perform poorly on the yeast dataset.
-For Minimum Sensitivity, C-MIEN4 obtains the best result in five out of seven datasets while ROAA outperforms it in both car and glass datasets.

For C-MIEN5, it is possible to conclude that:

For F-measure, it outperforms all of the algorithms in most datasets and is only outperformed by BROAA in one of them which indicates the algorithm's stronger performance on multi-class imbalanced datasets.
For G-Mean, again it outperforms most of the algorithms on the majority of the datasets but is outperformed by C-MIEN4 in two datasets and by BROAA in another one, meaning that either the ensemble variation or the normal one should get a better prediction for the minority class.
For Minimum Sensitivity, even though it outperforms all of the state of the art methods, it cannot improve the classification process over C-MIEN4 since both obtained the same result.

That said, C-MIEN's variations mitigate class imbalance and in general improve the performance of classifications better than traditional and state of the art decomposition methods.

4.1.2 Feature Selection

Feature selection is the process of selecting a subset of relevant features that allow a classifier to achieve better performance and improve its accuracy. This process can be divided into three distinct methods: filters, wrappers and embedded.[GE03] When searching for techniques to deal with imbalanced data, its possible to understand that there is significantly less research being done on feature selection methods than sampling techniques and that most techniques applied tend to convert multi-class into binary problems using class decomposition. In the following chapter an

attempt to identify techniques that deal specifically with multi-class without decomposition is made.

IAFN-FS

A wrapper feature selection method designed to solve multi class imbalance problems directly applied to multi-class or through means of class decomposition, Incremental ANOVA and Functional Networks-Feature Selection (IAFN-FS)[SMABCE09] uses sensitivity analysis on the AFN algorithm[Cas98] to find which subset, from all the possible subsets of features, performs better on the induction algorithm. This process has a low computation time and allows to discard irrelevant features while also taking into account interactions between features. Also, since the main indicative of a features' process of selection is variance, the results can be studied to understand why a feature is being rejected or selected which can prove helpful in different fields. Although the comparative results tend to point to better accuracy in class decomposition using this technique compared to a direct application to multi-class scenarios, the total number of selected features is bigger. This may be the result of a generalization while selecting each class' most important features.

FtCBF & FCCF

Fast Targeted Correlation-Based Filter (FtCBF) and Fast Class Correlation Filter (FCCF) are two scalable multi-class feature selection techniques proposed by Chidlovskii et al.[CL08] based on the Fast Correlation-Based Filter (FCBF) method which is widely regarded as one of the best state of the art algorithms[LY05] for feature selection on binary problems. These techniques attempt to improve upon the poor performance of its basis algorithm when dealing with more than two classes. The problem lies in FCBF's tendency to discard features it considers redundant when it is not the case, due to flaws when approximating the Markov blankets. "*Let we dispose a dataset S with feature set F and class set Y. A relevant feature Fi* $\subset$ *F is redundant if it has a Markov blanket in F where a Markov blanket for feature Fi is a feature subset Mi* $\subset$ *F which subsumes the information feature Fi has about target Y and all other features:*"

$$P(F - \mathcal{M}_i - \{F_i\}, Y|F_i, \mathcal{M}_i) = P(F - M_i - \{F_i\}, Y|\mathcal{M}_i).$$

The algorithm also uses the goodness criterion (GOOD), which classifies a feature as good if its weighted relevance is greater than the defined value, in order to greedily select which are relevant and which are redundant eliminating the second ones. FtCBF modifies its predecessor by adapting it to multi-class and adding extra conditions to the GOOD criterion with the objective of avoiding the issue. This helps to filter possible redundant features while improving its accuracy. On the other hand FCCF tackles the problem directly at the root by analyzing each class' uncertainty and feature correlation and taking them into consideration when approximating the Markov blanket. Although both techniques' purpose is not to deal with class imbalance specifically, they

also thrive under it and outperform the binary algorithm they are based on, FCBF, in several datasets. The downside of all the methods is the lack of consideration for feature interactions.

4.2 Algorithm-level approaches

Algorithm-level methods attempt to identify flaws in the classifier training procedure in order to handle imbalance data, shifting the focus followed by data-level solutions of combating class skewness in the training dataset. This is achieved by modifying the classifier in order to better deal with uneven data. The entire process requires in-depth comprehension about the classifier's mechanics in order to understand what might be the reason for the majority class to be biased. The main advantage compared to sampling solutions is that there are no changes to the dataset distribution which means it will work for various types of imbalance with the tradeoff being the specificity of the classifier type.[FGG+18] When it comes to multi-class imbalance data, there is an even higher cost associated which is the need to develop new algorithms since the difficulties of imbalance are more prevalent in this scenario. That said, most of the research invested into algorithm-level solutions is associated with the combination between different algorithms and ensemble learning or cost-sensitive frameworks since better results are yielded by using various classifiers. Pure algorithm modifications for multi-class imbalance are less common and usually revolve around decision tree adaptations.

Modified multi-class HDDT

Decision trees consist of some of the most fundamental used algorithms in the field of data mining, the most notorious being C4.5. An alternative method for building decision trees which was converted to deal with multi-class is Hellinger Distance Decision Trees (HDDT). MC-HDDT[HQCZ12] attempts to overcome its predecessor's inability to deal with more than two classes by improving the split criteria and reducing multiple classes into all possibilities of binary classes in an approach similar to class decomposition. Given a set of multiple classes C, identify each unique pair of subsets by: $C1 \subset C$, C2 = $\frac{C}{C1}$. After this all classes in C1 will be elected as the positive class while all the classes in C2 will act as the negative class. Then the Hellinger distance splitting criterion will find the best split between all choices of the negative and positive classes. From the results obtained and in comparison with a combination of both OAA and HDDT in an ensemble approach, it is concluded that MC-HDDT performs better in multi-class scenarios but is outperformed in binary problems, a situation where it is strongly skew insensitive.

4.3 Cost-sensitive learning

Another approach to dealing with data imbalance, in either a binary or multi-class scenario, is assuming heavier costs when misclassifying minority class examples. This process can be applied both at data-level, by defining costs during re-sample or feature selection, and at algorithmic-level, by changing the algorithm to be sensitive to the cost of minority class. Although a less popular method than, for instance, resampling, a comparison between cost-sensitivity and both data-level or algorithm-level approaches points to better computational resource usage. The downside is the difficult and time-consuming process of defining an efficient cost matrix to represent said misclassification. Also, it should require expertise on the topic, since cost attribution should be a thoroughly thought procedure. Research on misclassification costs points to a categorization between two types, example-dependent costs and class-dependent costs.[ZL06] The first one assumes that each example should have a misclassification cost while the second assumes that the bad classification should be applied to each of the classes. As expected, the former option is only followed in specific situations where example cost classification is done with ease, while the latter is more feasible in any scenario. Thus said, cost-sensitive learning can also be combined with boosting algorithms to produce ensemble techniques as an attempt to yield better results.

Rescale-new

Rescale-new [ZL06] was designed with the purpose of adapting the common rebalance of classes approach used in cost-sensitive learning to tackle multi-class problems while also demonstrating good results on imbalanced data. First, for a given cost matrix, an associated coefficient matrix is generated. If the rank of the coefficient matrix is smaller than the number of classes (consistent cost matrix), a non trivial solution must be found from the coefficient matrix. This is only possible if the value of matrix determinant is 0. The value of the non trivial solution will then be computed and used to rescale all the classes. This rescaled data set is then given to any cost-blind classifier; In the scenario that the value of the determinant is different than 0, the cost matrix is called inconsistent and, inevitably, the multi-class problem must be decomposed into pairs of two-class problems, where each two-class dataset is rescaled and given to a cost-blind classifier. The final prediction is made by voting the class labels predicted by the two-class classifier similarly to a class decomposition approach. From the compared results with traditional rescaling approaches such as oversampling and undersampling techniques (which often fail to reduce total cost on multi-class problems) and cost-blind learners, it is noticeable that the Rescale-New shows a significant improvement on multi-class problems. It does however extremely affected by severe class-imbalance, being outperformed by said approaches. It is also concluded that rescaling directly on multi-class problems can only obtain good performance when the costs are consistent and, as such, an examination of the cost consistency should be taken prior to applying this technique. Otherwise rescaling should be attempted after decomposing the multi-class problem into a compound of binary problems.

4.4 Ensemble-based approaches

Ensemble-based approaches take advantage of all the techniques discussed before to build stronger classifiers to deal with imbalanced problems. Due to overall producing the best results in classifying skewed data, it is common for the most used algorithms to be extended to deal with the multi-class scenario. Thus said, ensemble algorithms can be divided into three subcategories which are, Bagging, Boosting and Hybrid methods.

4.4.1 Bagging

SMOTEBagging

A SMOTE adaptation specifically developed to deal with multi-class imbalanced.[WY09] Unlike other binary techniques such as UnderBagging and OverBagging, class distribution is taken into consideration among all minority classes after sampling instead of over-sampling each class independently by using a different N value. A SMOTE resampling rate (b) will be set for each iteration and it will range from 10% to 100% while always being a multiple of 10. This ratio will define the number of minority class instances (b $\cdot Nm_{aj}$) which will be resampled with replacement from the original dataset in each iteration. Meanwhile, the SMOTE algorithm will generate the rest of the synthetic minority class instances. The results achieved in comparison with UnderBagging and OverBagging show that this method will outperform both binary and multi-class imbalance scenarios.

Multi-class Roughly Balanced Bagging

Created as an extension to Roughly Balanced Bagging which doesn't decompose multi-class problems into binary, MRBB[LS18] is a technique that attempts to learn all classes simultaneously. In this method, the main modification concerns the construction of bootstrap samples. While in the original approach the number of majority examples is decided by an estimation according to a negative binomial distribution, now a multinomial distribution is considered by using the classes' prior probabilities. This way, minority class instances will be more likely to be replicated than majority ones within each bag. Two versions of this model are also developed, and will be applied depending on the size of the bags: oMRBBag which considers that each bag contains the same number of instances than the original dataset and uMRBBag which will limit the size of the bags to be equal to that of the minority class.

SOUP-Bagging

Similarity Oversampling and Undersampling Preprocessing (SOUP)[JLS19] is a technique that utilizes hybrid sampling in order to obtain a more balanced distribution of classes in datasets. This method works by resampling the original dataset iteratively and with replacement, then applying the SOUP technique and finally by constructing a classifier. Stratified sampling is used as the

method of resampling. Lastly, a majority voting of the component classifiers is computed in order to obtain a prediction.[LS]

4.4.2 Boosting

Adaboost.NC for multi-class

An extension developed to handle multi-class scenarios was proposed to Adaboost.NC[WY12b] to obtain a better performance in said scenario without using class decomposition. This technique adds negative correlation training to boosting and then, by combining it with oversampling minority class examples can be identified more easily which contributes to the overall better predictions when handling more than two classes. When analyzing the performance of Adaboost.NC the results show that there is no improvement in using class decomposition for multi-class imbalanced learning. In fact, it even lowers the performance due to loss of important information.

BPSO-AdaBoost-KNN

This method is a combination between a feature selection technique, BPSO,[CXZS12] which is applied to search for the best subset of all features, a boosting technique, AdaBoost, which is an effective tool to improve the learning ability of classifiers by integrating them into a stronger one and widely regarded as the most effective ensemble method, and a base classifier, KNN, which although being very simple as an effective classifier was chosen due to its low computational cost and by previously achieving good results in multi-class scenarios.
Firstly BPSO is implemented as the feature selection algorithm and then the conjunction classifier Adaboost-KNN converts the traditional baseline classifier KNN into a strong classifier. Finally, both of these two processes are measured by a new evaluation metric called AUCarea in order to pick the best feature subset as well as the final prediction of the ensemble model thus evaluating the performance of the BPSO-Adaboost-KNN technique.[HYY+16]

4.4.3 Hybrid methods

HECMI

Hybrid Ensemble for Classification of Multiclass Imbalanced (HECMI)[STBM17] is a technique which combines both data and algorithm based approaches to handle the scenario of multi-class imbalance, to create an hybrid of boosting and bagging which focuses on the misclassified instances while also applying data splitting techniques and obtaining predictions by majority voting. The algorithm works by creating iteratively a diverse ensemble of classifiers which are maintained by the way the data is sampled while creating the model. The number of classifiers in the ensemble is proportional to the number of classes which are not known a priori. HECMI performs at par with the traditional classification techniques when the imbalance ratio is small to moderate. However, for datasets with high imbalance ratio, HECMI has proven to give better recall rates for minority classes.

4.5 Evaluation Metrics

As far as evaluation metrics for imbalanced data are concerned, an in-depth survey was published by Torgo *et al.*[BTR16] that addresses performance metrics in different scenarios. When dealing with multi-class imbalance the usage of common metrics such as accuracy is not advised, since there is no clear distinction between classification values on different classes. This bias may lead to inaccurate conclusions.[FLG+13] There is a need for metrics that consider each class' individual performance when dealing with class imbalance and, as such, several metrics were purposed, the most notable being F-Measure, Geometric Mean (G-Mean) and AUC_{Roc}.

F-measure

This metric is possibly the most used to assess the performance of an algorithm when dealing with imbalanced scenarios. It represents the weighted average of the precision and recall metrics while taking into account for false positives (FP) and for false negatives (FN) and is given by the following equations:

$$PRecision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$

$$F-Measure = 2 * \frac{Recall * Precision}{Recall + Precision}$$

G-Mean

Geometric mean(G-Mean) is less sensitive to value skewness and especially outlier presence than the commonly used arithmetic mean. It evaluates the performance of a given classifier when predicting positive examples, something that should be taken into account since that even if negative examples are being badly classified, an algorithm can still present a good G-Mean value. It is given by the following equation:

$$G\text{-}mean = \sqrt{TPR \cdot TNR}$$

AUC_{Roc}

AUC_{Roc} curves are performance measurements used for classification problems that attempt to describe a model's ability to distinguish between classes. ROC is a probability curve and AUC

represents the degree or measure of separability. An higher value of AUC, means that the model is better at predicting true positives. It is given by the following plot between true positive and false positive ratios [auc19]:

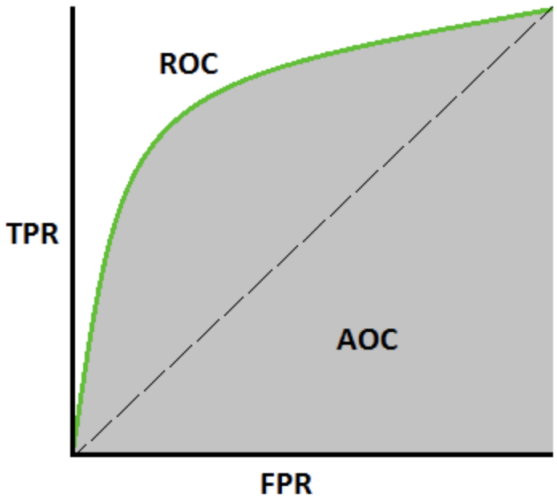


Figure 4.1: AUC_{Roc} curve plot[auc19]

4.6 Conclusions on the survey

An attempt of using M.Galar *et al.* taxonomy was made to categorize algorithms that do not use class decomposition in order to handle class imbalance. This survey started by explaining decomposition schemes to understand its strengths and downsides and then was followed by analysing algorithms from the four described approaches while giving insight into the authors' results when compared to other techniques relevant to the discussion. Lastly, it was wrapped up with an explanation on the evaluation metrics that should be used under this scenario. The result was positive since it was possible to find at least one method that could fit into all the categories, even though further research should be considered since, with the rising interest on the topic, new techniques or even new taxonomies may be developed. It was also possible to understand that sampling techniques and ensemble based methods are the most frequently used and often yield better performances compared to the remaining approaches. This conclusion also motivated the next chapter where some of the discussed ensemble methods will be benchmarked and compared.

Chapter 5

Techniques benchmark and analysis of results

In this section we will be comparing two of the previously discussed ensemble based techniques, MRBB and SOUPBagging.

5.1 Reasoning and Experimental Setup

The choice of the algorithms was based upon being, in their respective approaches, the most recently published methods discussed in the survey and as such, having better performances comparatively to older techniques. Also, the authors' results for each of the methods do not show comparison between them.

The experimental setup consists of 9 datasets that present different imbalance ratios and can be found on UCI Machine Learning repository. The models were developed trained and tested with the help of the multi-imbalance python package, which is based on scikit-learn and makes available implementations of some state of the art multi-class imbalanced algorithms. The classifiers chosen for the ensemble techniques was the KNN due to being simple yet effective on multiclass imbalanced scenarios and Decision Tress. For evaluation of performance, the G-Mean metric was used. The dataset information is displayed in the following table, where IR represents the imbalance ratio which was calculated by diving the majority class instances for the minority class' and where IC represents the type of imbalance category.

No.	Dataset	Instances	Attributes	Classes	Class Distribution	IR	IC
1	new-thyroid	215	5	3	150/35/30	5	multi-majority
2	yeast	1484	8	10	463/429/244/163/51/44/37/30/20/5	92.6	mixed
3	car	1728	6	4	1210/384/69/65	18.62	multi-majority
4	cleveland	303	14	5	164/55/36/35/13	12.62	multi-minority
5	balance-scale	625	4	3	288/49/288	5.88	multi-majority
6	dermatology	366	33	6	112/61/72/49/52/20	5.6	multi-minority
7	winequality-red	1599	12	6	10/53/681/638/199/18	68.1	mixed
8	ecoli	336	7	8	143/77/52/35/20/5/2/2	76.5	multi-minority
9	cmc	443	7	3	629/333/51	12.33	multi-majority

Table 5.1: Summary of dataset information.

5.2 Results

5.2.1 Ensemble-Based Algorithms

Dataset	MRBB-KNN	MRBB-DT	SOUPBag-KNN	SOUPBag-DT
new-thyroid	0.730	**0.977**	0.897	0.953
yeast	0.479	**0.496**	0.480	0.485
car	**0.957**	0.811	0.851	0.953
cleveland	0.312	**0.327**	0.061	0.105
balance-scale	0.704	0.637	**0.750**	0.648
dermatology	0.722	**0.980**	0.802	0.957
winequality-red	0.361	**0.531**	0.385	0.470
ecoli	**0.808**	0.774	0.806	0.735
cmc	**0.545**	0.531	0.485	0.498

Table 5.2: Average G-Mean of the ensemble algorithms.

From the results obtained by running the experiments five times using sklearn train_test_split module for input data shuffling, we can visualize the following :
- *1st)* Overall performance is similar in both algorithms across all the datasets except for cleveland, but MRBB using any of chosen classifiers outperforms SOUPBag in 8 out of 9 datasets.
- *2nd)* All algorithms perform poorly in the yeast, winequality-red and cmc and extremely poorly in cleveland datasets.
- *3rd)* Using Decision Trees as a classifier seems to be performing better for MRBB, with MRBB-DT outperforming its counterpart in 5 out of 9 datasets while KNN is a better choice for SOUP, with SOUPBag-KNN outperforming its counterpart in 7 out of 9 datasets.
- *4rd)* The performance of both algorithms does not seem to be affected by the type of imbalance category, although they do perform poorly in mixed scenarios.

While the first assessment does not give us much information besides that MRBB seems to be a better method comparatively with SOUPBagging, the second assessment gives us an important insight about the liability of both algorithms to extreme outlier datasets that contain rare classes (contain fewer than 5 examples) such as cleveland. This effect was already described when the methods were purposed. The third assessment is linked with the implementation of the algorithms itself, since most extensions of rough bagging algorithms are usually constructed using decision

trees. There is no reported reason for the fourth assessment, although it may be linked to the explanation given about the second insight and the dataset structure itself, which means further testing should be conducted.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

From the main objectives defined at the start of this study, a taxonomy to categorize algorithms that deal with class skewness was adapted to accommodate techniques that deal with multi-class imbalance without class decomposition. A survey of existing algorithms was conducted which lead to the conclusion that ensemble methods and resampling, especially oversampling, should be go-to techniques since they are the best performers for this scenario. From this conclusion a benchmark and comparison between the two latest state of the art ensemble methods developed specifically for multi-class imbalance was conducted.
The results were positive and allowed us to assess strengths and flaws on both algorithms and understand which classifiers should be used for different scenarios. Overall, the balance is positive since, the main objectives proposed were met.

6.2 Future Work

From the study made it was possible to underline some key aspects to be investigated further, such as, the need for a new report based on comparison between all of the algorithms described in the different approaches in the same experimental setting and applied to the maximum possible amount of datasets. This would allow inexperienced users in the field to produce results since they already knew what algorithm to use based on the problem.
Finding and discussing methods and approaches outside of the proposed taxonomy, in areas such as hierarchical learning,[AS17] not only could diversify it, but also might prove valuable since little research is done but results look promising.
Lastly, there is always room for the proposal of a new method, either by extending a binary specific algorithm to multi-class or by developing a new directly method.

References

[AAZ+18] Tahira Alam, Chowdhury Farhan Ahmed, Sabit Anwar Zahin, Muhammad Asif Hossain Khan, and Maliha Tashfia Islam. An effective ensemble method for multi-class classification and regression for imbalanced data. In *Industrial Conference on Data Mining*, pages 59–74. Springer, 2018.

[AH15] Lida Abdi and Sattar Hashemi. To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering*, 28(1):238–251, 2015.

[AS17] Hanaa Abdalaziz and Fakhreldeen Saeed. New hierarchical model for multiclass imbalanced classification. *Journal of Theoretical and Applied Information Technology*, 3195, 08 2017.

[auc19] Auc roc curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5, 2019.

[AVP15] Astha Agrawal, Herna L. Viktor, and Eric Paquet. Scut: Multi-class imbalanced data classification using smote and cluster-based undersampling. *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, 01:226–234, 2015.

[BFH06] Klaus Brinker, Johannes Fürnkranz, and Eyke Hüllermeier. A unified model for multilabel classification and ranking. volume 141, pages 489–493, 01 2006.

[bin19] Binary vs multiclass image. https://www.holehouse.org/mlclass/06_Logistic_Regression.html, 2019.

[BSL09] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 475–482. Springer, 2009.

[BTR16] Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2), August 2016.

[BZ18] Jingjun Bi and Chongsheng Zhang. An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *Knowledge-Based Systems*, 158, 06 2018.

[Cas98] Enrique Castillo. Functional networks. *Neural Process. Lett.*, 7(3):151–159, June 1998.

[CB91] Peter Clark and Robin Boswell. Rule induction with cn2: Some recent improvements. In Yves Kodratoff, editor, *Machine Learning — EWSL-91*, pages 151–163, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[CBH] Nitesh V Chawla, Kevin W Bowyer, and Lawrence O Hall. Kegelmeyer. wp (2000). *SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research*, 16(1):321–357.

[CKSLS14] Marie Chavent, Vanessa Kuentz-Simonet, Amaury Labenne, and Jérôme Saracco. Multivariate analysis of mixed data: The r package pcamixdata. *arXiv preprint arXiv:1411.4911*, 2014.

[CL08] Boris Chidlovskii and Loïc Lecerf. Scalable feature selection for multi-class problems. volume 5211, pages 227–240, 09 2008.

[CN89] Peter Clark and Tim Niblett. The cn2 induction algorithm. *Machine learning*, 3(4):261–283, 1989.

[CSSC18] Rafael MO Cruz, Mariana A Souza, Robert Sabourin, and George DC Cavalcanti. Icprai 2018 si: On dynamic ensemble selection and data preprocessing for multi-class imbalance learning. *arXiv preprint arXiv:1811.10481*, 2018.

[CXZS12] Liam Cervante, Bing Xue, Mengjie Zhang, and Lin Shang. Binary particle swarm optimisation for feature selection: A filter based approach. In *2012 IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE, 2012.

[DWT+18] Hao Ding, Bin Wei, Ning Tang, Zhibin Yu, Nan Wang, Haiyong Zheng, and Bing Zheng. Plankton image classification via multi-class imbalanced learning. In *2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO)*, pages 1–6. IEEE, 2018.

[FGG+18] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. Algorithm-level approaches. In *Learning from Imbalanced Data Sets*, pages 123–146. Springer, 2018.

[FLG+13] Alberto FernáNdez, Victoria LóPez, Mikel Galar, MaríA José Del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42:97–110, 2013.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[GFB+11] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

[goo19] Imbalanced learning. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data, 2019.

REFERENCES

[GPOB06] Nicolas Garcia-Pedrajas and Domingo Ortiz-Boyer. Improving multiclass pattern recognition by the combination of two strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):1001–1006, 2006.

[HBGL08] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[HQCZ12] T Ryan Hoens, Qi Qian, Nitesh V Chawla, and Zhi-Hua Zhou. Building decision trees for the multi-class imbalance problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 122–134. Springer, 2012.

[HWM05] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[HYS+17] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[HYY+16] Guo Haixiang, Li Yijing, Li Yanan, Liu Xiao, and Li Jinling. Bpso-adaboost-knn ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*, 49:176–193, 2016.

[Irv92] AD Irving. Stochastic sensitivity analysis. *Applied mathematical modelling*, 16(1):3–15, 1992.

[JLS19] Małgorzata Janicka, Mateusz Lango, and Jerzy Stefanowski. Using information on class interrelations to improve classification of multiclass imbalanced data: A new resampling algorithm. *International Journal of Applied Mathematics and Computer Science*, 29(4):769–781, 2019.

[KKW17] Michał Koziarski, Bartosz Krawczyk, and Michał Woźniak. Radial-based approach to imbalanced data oversampling. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 318–327. Springer, 2017.

[KKW19] Bartosz Krawczyk, Michał Koziarski, and Michal Wozniak. Radial-based oversampling for multiclass imbalanced data classification. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–14, 06 2019.

[KMC17] Bartosz Krawczyk, Bridget Mcinnes, and Alberto Cano. Sentiment classification from multi-class imbalanced twitter data using binarization. 06 2017.

[Kov19] György Kovács. Smote-variants: A python implementation of 85 minority oversampling techniques. *Neurocomputing*, 366:352–354, 2019.

[Lan19] Mateusz Lango. Tackling the problem of class imbalance in multi-class sentiment classification: An experimental study. *Foundations of Computing and Decision Sciences*, 44(2):151 – 178, 2019.

[LDCG08] Ana Carolina Lorena, André CPLF De Carvalho, and João MP Gama. A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30(1-4):19, 2008.

[Lia08] T Warren Liao. Classification of weld flaws with imbalanced class data. *Expert Systems with Applications*, 35(3):1041–1052, 2008.

[LS] Mateusz Lango and Jerzy Stefanowski. Soup-bagging: a new approach for multi-class imbalanced data classification.

[LS18] Mateusz Lango and Jerzy Stefanowski. Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data. *Journal of Intelligent Information Systems*, 50(1):97–127, 2018.

[LY05] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[MWOF07] Yi L Murphey, Haoxing Wang, Guobin Ou, and Lee A Feldkamp. Oaho: an effective algorithm for multi-class learning from imbalanced data. In *2007 International Joint Conference on Neural Networks*, pages 406–411. IEEE, 2007.

[NHY+15] W. W. Y. Ng, J. Hu, D. S. Yeung, S. Yin, and F. Roli. Diversified sensitivity-based undersampling for imbalance classification problems. *IEEE Transactions on Cybernetics*, 45(11):2402–2412, Nov 2015.

[PA04] GEA PA. Batista. *RC Prati, MC Monard, A study of the behaviour of several methods for balancing machine learning training data, SIGKDD Explor*, 6(1):20–29, 2004.

[PMS18] Shekhar Pandey, Supriya M, and Abhilash Shrivastava. Data classification using machine learning approach. In Sabu M. Thampi, Sushmita Mitra, Jayanta Mukhopadhyay, Kuan-Ching Li, Alex Pappachen James, and Stefano Berretti, editors, *Intelligent Systems Technologies and Applications*, pages 112–122, Cham, 2018. Springer International Publishing.

[PPEP19] Isha Pradhan, Katerina Potika, Magdalini Eirinaki, and Petros Potikas. Exploratory data analysis and crime prediction for smart cities. In *Proceedings of the 23rd International Database Applications & Engineering Symposium*, pages 1–9, 2019.

[PRW16] Kamthorn Puntumapon, Thanawin Rakthamamon, and Kitsana Waiyamai. Cluster-based minority over-sampling for imbalanced datasets. *IEICE TRANSACTIONS on Information and Systems*, 99(12):3101–3109, 2016.

[PS12] Wanthanee Prachuabsupakij and Nuanwan Soonthornphisaj. Clustering and combined sampling approaches for multi-class imbalanced data classification. In *Advances in information technology and industry applications*, pages 717–724. Springer, 2012.

[PS14] Wanthanee Prachuabsupakij and Nuanwan Soonthornphisaj. Cluster-based sampling of multiclass imbalanced data. *Intelligent Data Analysis*, 18(6):1109–1135, 2014.

[S+14] TD Sravani et al. Multiclass unbalanced protein data classification using sequence features. In *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–8. IEEE, 2014.

REFERENCES

[SD18] Naman D. Singh and Abhinav Dhall. Clustering and learning from imbalanced data, 2018.

[SKW06] Yanmin Sun, Mohamed S Kamel, and Yang Wang. Boosting for learning multiple classes with imbalanced class distribution. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 592–602. IEEE, 2006.

[SMABCE09] Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, and Rosa M Calvo-Estévez. A wrapper method for feature selection in multiple classes datasets. In *International Work-Conference on Artificial Neural Networks*, pages 456–463. Springer, 2009.

[STBM17] Suresh Chandra Satapathy, Joao Manuel RS Tavares, Vikrant Bhateja, and JR Mohanty. Information and decision sciences. In *Proceedings of the 6th International Conference on FICTA*. Springer, 2017.

[sup19] Supervised vs unsupervised learning. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d, 2019.

[WM97] D Randall Wilson and Tony R Martinez. Improved heterogeneous distance functions. *Journal of artificial intelligence research*, 6:1–34, 1997.

[WY09] Shuo Wang and Xin Yao. Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 324–331. IEEE, 2009.

[WY12a] S. Wang and X. Yao. Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1119–1130, Aug 2012.

[WY12b] Shuo Wang and Xin Yao. Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1119–1130, 2012.

[YKZZ17] Xuebing Yang, Qiuming Kuang, Wensheng Zhang, and Guoping Zhang. Amdo: an over-sampling technique for multi-class imbalanced problems. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1672–1685, 2017.

[ZBX+19] Chongsheng Zhang, Jingjun Bi, Shixin Xu, Enislay Ramentol, Gaojuan Fan, Baojun Qiao, and Hamido Fujita. Multi-imbalance: An open-source software for multi-class imbalance learning. *Knowledge-Based Systems*, 174, 03 2019.

[ZL06] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, page 567–572. AAAI Press, 2006.